# Crowdsourcing without a crowd: Reliable online species identification using Bayesian models to minimize crowd size

ADVAITH SIDDHARTHAN, University of Aberdeen
CHRISTOPHER LAMBIN, University of Aberdeen
ANNE-MARIE ROBINSON, University of Aberdeen
NIRWAN SHARMA, University of Aberdeen
RICHARD COMONT, Bumblebee Conservation Trust
ELAINE O'MAHONY, Bumblebee Conservation Trust
CHRIS MELLISH, University of Aberdeen
RENÉ VAN DER WAL, University of Aberdeen

We present an incremental Bayesian model which resolves key issues of crowd size and data quality for consensus labelling. We evaluate our method using data collected from a real world citizen science program, BEEWATCH, which invites members of the public in the UK to classify (label) photographs of bumblebees as one of 22 possible species. The biological recording domain poses two key and hitherto unaddressed challenges for consensus models of crowdsourcing: (a) the large number of potential species makes classification difficult and (b) this is compounded by limited crowd availability, stemming from both the inherent difficulty of the task and the lack of relevant skills among the general public. We demonstrate that consensus labels can be reliably found in such circumstances with very small crowd sizes of around 3–5 users (i.e. through group sourcing). Our incremental Bayesian model, which minimizes crowd size by re-evaluating the quality of the consensus label following each species identification solicited from the crowd, is competitive with a Bayesian approach that uses a larger but fixed crowd size and outperforms majority voting. These results have important ecological applicability: biological recording programs such as BEEWATCH can sustain themselves when resources such as taxonomic experts to confirm identifications by photo submitters are scarce (as is typically the case), and feedback can be provided to submitters in a timely fashion. More generally, our model provides benefits to any crowdsourced consensus labeling task where there is a cost (financial or otherwise) associated with soliciting a label.

Categories and Subject Descriptors: 1.7 [**Systems and Applications**]: AI and environmental protection; 3.2 [**Methodology**]: Emerging applications and technology; 2.8 [**AI Technology**]: Machine Learning

Additional Key Words and Phrases: Crowdsourcing, Citizen Science, Consensus Model, Bayesian Reasoning, Bumblebee Identification, Biological Recording

## 1. INTRODUCTION

The term 'crowdsourcing' is often used in citizen science to refer to models of data collection or annotation that involve the general public, so that initiatives can be scaled up beyond what a small number of experts could achieve among themselves.

Author's addresses: Advaith Siddharthan, Nirwan Sharma and Chris Mellish, Computing Science, University of Aberdeen, Aberdeen AB24 3UE, UK, emails: {advaith, n.sharma, c.mellish}@abdn.ac.uk; Christopher Lambin, Anne-Marie Robinson and René van der Wal, Aberdeen Centre for Environmental Sustainability (ACES), University of Aberdeen, Aberdeen AB24 3UU, UK, emails: christopher.lambin@gmail.com, annierobinson@abdn.ac.uk, r.vanderwal@abdn.ac.uk; Richard Comont and Elaine O'Mahony, Bumblebee Conservation Trust, Cottrell Building, University of Stirling, Stirling FK9 4LA, UK, emails: richardcomont@bumblebeeconservation.org, elaineomahony@bumblebeeconservation.org.

DOI:http://dx.doi.org/10.1145/0000000.0000000

In the biological recording domain, key success stories tend to come from ornithology, a scene that can rely on a large number of observers with high skills and strong self-motivation [Greenwood 2007]. For example, Cornell University's eBird has become a huge volunteer-based biological data gathering program with bird records now being submitted from all over North America and beyond [Hochachka et al. 2012]. Given the importance of data reliability, various data validation routines were developed for this large program, including those where an expert could request additional information about sightings interactively to confirm unusual records [Bonter and Cooper 2012].

An increasingly common method to improve data reliability in crowdsourcing is to frame the exercise as a consensus task, with the goal to identify a hidden state of the world by aggregating assessments from multiple participants [Kamar et al. 2012]. Arguably the most successful example of this is Galaxy Zoo [Lintott et al. 2008], where amateur astronomy groups worldwide classified galaxies in photos taken by the Hubble telescope as either spiral or elliptical. Advances in digital photography and widespread societal adoption thereof have led to a rapidly growing interest in using the consensus model for biological recording, where volunteer recorders (or even camera traps) upload photographs taken of specimens to have these subsequently identified by a crowd of other volunteers. For example, the Snapshot Serengeti (www.snapshotserengeti.org) project invites lay people to identify large mammals from photographs taken with camera traps from across this biodiversity hotspot.

Applications of crowdsourcing in consensus tasks typically assume (a) the availability of a large crowd, and (b) a relatively straightforward classification task. When these assumptions hold, simple voting models (such as majority vote) can be used to combine crowd assessments; that way large amounts of data can be annotated with limited involvement of experts. Where these assumptions do not hold, however, there is a need for validated methods which minimize the required crowd size and estimate certainties of consensus identifications. This is particularly pertinent when there are time or cost constraints; for instance, obligations to provide prompt feedback to or financially compensate those who have submitted data or annotations.

In this article we make the following methodological, resource and applied contributions:

(1) We present an incremental formulation of a Bayesian consensus model and demonstrate that it is as accurate as a Multinomial Naive Bayes model that uses a fixed crowd size. Our method requires much smaller crowd sizes, which provides clear benefits to any crowdsourced consensus labeling task where there is a cost (financial or otherwise) associated with soliciting a crowd label.
(2) We present, and make available for research, a novel dataset collected from a real world citizen science program in the biological recording realm, BEEWATCH, which is challenging in two key respects. First, there is a large set of labels (22 possible species of bumblebee), and second, the task is relatively difficult for humans, which leads to a high level of noise (average accuracy of a crowdsourced label is 59%). The dataset includes expert labels to facilitate supervised machine learning.
(3) We show that reliable consensus identifications can be achieved for this dataset through employing incremental Bayesian methods, requiring very small crowd sizes of around 3–5 users (i.e. group sourcing). This has important ecological applicability: biological recording programs such as BEEWATCH can sustain themselves without major additional resources (including taxonomic experts to confirm identifications, who have become increasingly scarce), and feedback can be provided to submitters in a timely fashion.

## 2. RELATED WORK

Empirical aspects of crowdsourcing models have only recently emerged as topics of investigation. Studies have shown that increasing crowd size results in improved accuracy of the consensus label for both Majority Voting and Bayesian Models [Sheng et al. 2008; Loni et al. 2014]. Building on these, several other papers have focused on identifying the subset of a crowd with the best skills for the task at hand [Li et al. 2014; Karger et al. 2014], often in the context of active learning, a semi-supervised machine learning approach where labels are sought selectively for the examples most likely to boost machine learning performance [Donmez and Carbonell 2008; Yan et al. 2011; Ipeirotis et al. 2014]. Others have modeled crowd member retention to predict when a worker would disengage from their assigned tasks [Mao et al. 2013]. For biological recording, iSpot (www.ispot.org.uk) modeled the *reputation* of users [Clow and Makriyannis 2011; Silvertown et al. 2015], taking into account both the activity of a user (e.g. numbers of records posted and identifications made) and their accuracy (e.g. agreement with expert users). While these studies show that repeated labeling and soliciting better skilled workers improves accuracy, they do not address the problem of minimizing crowd size.

Indeed, published work on the aggregation of crowd labels have typically used a predetermined crowd size. For instance, the TurKit toolkit [Little et al. 2009] for creating and managing tasks in Amazon's Mechanical Turk provides an implementation of a voting function for binary classification. It recruits workers until the number of votes for one of the (two) options is greater than a specified threshold (e.g. 8/10). Other models have been developed for binary classification; for instance, GLAD (Generative model of Labels, Abilities, and Difficulties) was developed for classifying an image of a face as smiling or not smiling [Whitehill et al. 2009]. GLAD simultaneously infers the expertise of each 'labeler', the difficulty of each image and the most probable label for each image. However, since it performs joint estimation of consensus labels and model parameters, it can only be run after the completion of the program, and consequently (a) it cannot be used to determine crowd size or consensus label likelihood incrementally, and (b) recorders cannot be provided with feedback on their submission until the end of the program.

In this article we present a model which addresses the key crowdsourcing constraints that arise when using consensus models for biological recording. Our approach integrates new information (a crowdsourced species identification) in an incremental fashion, thereby allowing a recalculation of the likelihood of the consensus being correct every time an identification is submitted. This is used to inform the solicitation of additional identifications and facilitates the delivery of prompt feedback to the photo submitter where required. We base our work on the Naive Bayesian family of methods which have been shown to be effective for modeling and correcting bias in consensus labeling tasks [Dawid and Skene 1979; Sheshadri and Lease 2013; Snow et al. 2008]. The novelty of the work presented here is the iterative application of the Bayesian method (i.e. the re-evaluation of the quality of the crowd consensus after each solicited label in order to minimize crowd size) and the demonstration of the effectiveness of the method for biological recording, which typically involves a large numbers of labels (22 species of bumblebees in this work).

## 3. MATERIALS AND METHODS

### 3.1. BEEWATCH as a test platform

We conducted our research through developing, with the Bumblebee Conservation Trust (BBCT; www.bumblebeeconservation.org), an online photo submission and identification platform called BEEWATCH (www.abdn.ac.uk/research/beewatch).

BEEWATCH allows members of the general public (henceforth, recorders) to submit photos of bumblebees, along with a location and date of sighting (i.e. the basic information required for a biological record). The recorder is then encouraged to identify the specimen in the photograph as one of the 22 species of bumblebee present in the UK[1] using an online identification guide. Through the interface shown in Fig. 1 (a), the recorder can select visual features of the bumblebee (e.g. color patterns on thorax and abdomen; shape of head and tail) to narrow down the possible species, select a species identification and then submit the record.

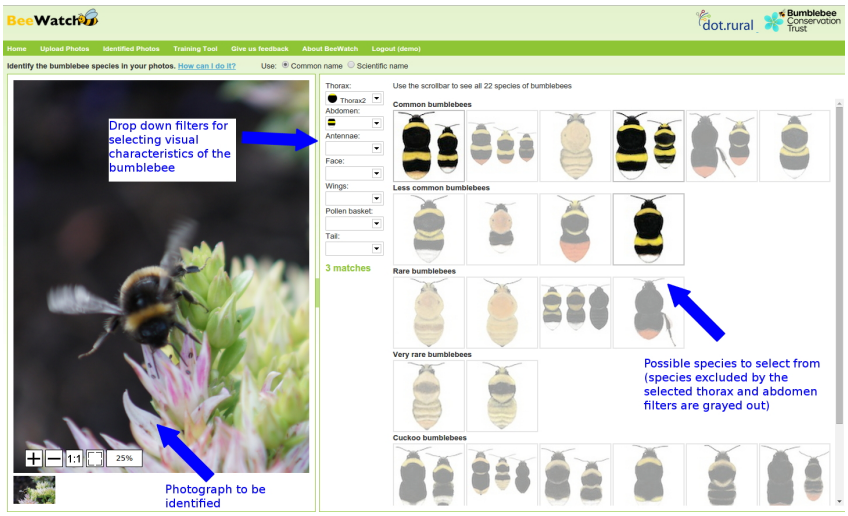
To ensure data quality, each submitted photo record is verified by a taxonomic expert at either the BBCT or the University of Aberdeen. These experts communicate the correct identification to the recorder by email, along with textual feedback aimed at helping the recorder improve their identification skills. In previous work [Blake et al. 2012], we described how the provision of this feedback could be automated using Natural Language Generation technology, through implementing influential ideas on formative feedback in learning. Automating the provision of feedback to recorders allowed BEEWATCH to scale up as a recording program, from handling 200 photo records in 2011 to over 4,000 in 2014 and also led to an improvement in the identification skills of citizen scientists [Blake et al. 2012]. By December 2014, BEEWATCH had collected over 10,000 verified photo records of bumblebees from across the UK.

Further scaling up of BEEWATCH is prevented by limited availability of taxonomic experts to review each photo record. Addressing this bottleneck – common to many biological recording programs – has been the prime motivation for developing a consensus model through which a large proportion of photo-records can be verified using other BEEWATCH users. This required BEEWATCH to be extended such that users could also provide an identification for photo-records submitted by others. Figure 1 (b) shows the interface of this 'consensus extension' through which users, after providing an identification for a photo-record, can see how others have identified the same photo-record. The system was set up such that new photo submissions were automatically sent to the consensus extension, where they were allowed to each accumulate up to 10 crowdsourced identifications before being replaced by a new photo submission. Recorders could only identify their submission as one of the 22 bumblebee species (as these are the focus of the recording scheme), but both expert identifiers and those users identifying photo-records of others through the consensus module had access to two additional labels: 'not a bumblebee' and 'not identifiable', in order to categorize the entirety of submissions.

### 3.2. The BEEWATCH consensus dataset

Through the consensus extension to BEEWATCH, and including the identification provided by the photo submitter when available, we collected in total 8,844 independent identifications by 763 users of 1,613 photo submissions between 5 May 2013 and 24 May 2014. This amounted to 5.48 (8,844/1,613) independent identifications per photo on average. To obtain data for this supervised learning study, we only chose photographs for which an expert identification existed, and left photo submissions on the consensus extension until they accumulated 10 independent identifications; however, given that there were on average only 5.48 identifications per photo submission, only

[1]There are 25 species of bumblebee in the United Kingdom, but three of these (*Bombus lucorum* L. *sensu strictu*, *Bombus cryptarum* Fabricius and *Bombus magnus* Vogt) cannot be reliably distinguished from each other based on visual characteristics alone. These form a species complex which, for the purposes of BEEWATCH, is treated as one species (*Bombus lucorum sensu lato*, the white-tailed bumblebee). The extinct species *Bombus subterraneus* L., the short-haired bumblebee, is also excluded despite an ongoing reintroduction attempt, because of the extremely low likelihood of recording it.




(a) Species identification interface.


(b) Post-identification feedback interface.

Fig. 1. Screenshots from the BEEWATCH biological recording web interface: (a) Interface for identifying a specimen in a photograph as one of 22 species of bumblebee, and (b) Interface to provide feedback to user following their identification of a photograph using the consensus module, by tabulating identifications by other users.

594 of the possible 1,613 managed to accumulate 10 crowd identifications. This data collection exercise already indicates some of the key challenges for applying crowdsourcing methods for biological recording:

(1) *Crowd availability:* Though the pool of users was substantial (763 different individuals contributed identifications to the study), they on average generated only 5.48 crowdsourced identifications per submitted photo. This highlights the requirement to work with as small as possible crowd sizes (i.e. group sourcing).
(2) *Differing user engagement and skills:* Of the 763 users, 314 only contributed a single identification each, 130 users contributed at least 10 identifications, 83 users contributed at least 20, 40 users contributed at least 50, and 14 contributed at least

100. Thus, whilst there were on average 11.6 identifications (of different photos) per user, there were large differences in the level of engagement with our crowdsourcing tool. The accuracy of individual identifications (averaged over all users and photos) in the data set was 59.2%. Among users who had identified at least 10 photos, user accuracy ranged between 18% and 90%.

(3) *Differing difficulty by species:* Accuracy also varied by species, from 22% for *Bombus jonellus* Kirby (heath bumblebee), which is similar in appearance to more common species such as *Bombus hortorum* L. (garden bumblebee), to 86% for *Bombus hypnorum* L. (tree bumblebee), a common and distinctive species. Figure 5 (a) shows how the species identifications by BEEWATCH users related to those by our experts.

The goal of this paper is to develop and evaluate an *efficient* consensus model for combining independent identifications of the same photo-record by different users, which *minimizes* crowd size by taking into account characteristics of both users and species. While in general crowdsourcing is used to solicit labels from non-experts as an inexpensive alternative to recruiting experts, in practice limited amounts of expert labels are often available to facilitate supervised learning [Tang and Lease 2011; Sheshadri and Lease 2013]. This is also the case for biological recording, and the dataset described here includes expert labels. Our model, described next, is therefore fully supervised.

### 3.3. Incremental Bayesian models for evaluating consensus

The Bayesian framework provides a straightforward means of using new evidence (in our case, a new identification by a user of a photo submitted by another user) to update an existing estimate of the likelihood of a *hypothesis* (also sometimes referred to as a *proposition*) being correct; in our case, a hypothesis is a possible species identity compatible with the collective identifications of the crowd so far.

The Bayesian framework is particularly well suited to a classification task with large numbers of categories (22 bumblebee species in our case). Intuitively, the likelihood of multiple users selecting the same species by chance is very low; therefore, when independent identifications of a specimen in a photo agree, there is a strong likelihood that this consensus identification is correct. The Bayesian framework gives us a means to directly estimate the likelihood that a photographed specimen is of a certain identity (the hypothesis, $H$) given the independent identifications by users (the evidence, $E_i$). We model two components of the evidence: (i) the ease of identification of a species (as some species are visually more distinctive than others, and thus easier to identify); and (ii) the identification skills of a user (as some users are better at the task than others).

We first present a model that only takes into account the ease of identification of a species (hereafter coined 'species model') and subsequently extend this by including user identification skill level (i.e 'user+species model') to further improve our ability to derive an accurate species identification.

Consider Bayes Rule in Odds Notation (see Appendix A for a full explanation of the odds notation and its derivation from the definitions of joint and conditional probabilities):

$$O(H|E_1, ..., E_n) = O(H) \times \Lambda(H|E_1) \times ... \times \Lambda(H|E_n) \tag{1}$$

$$\text{where, } \Lambda(H|E_i) = \frac{P(E_i|H)}{P(E_i|\neg H)} \tag{2}$$

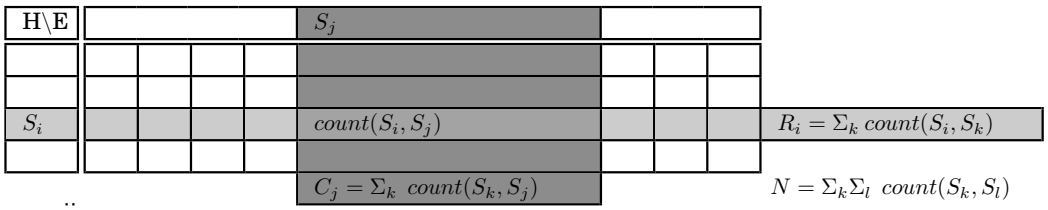



Fig. 2. Schematic representation of Fig. 5 (a), showing the counts used in the estimation of prior odds and $\Lambda$ terms from data. Each cell $count(S_i, S_j)$ is the count of how frequently a photo record identified as species $S_i$ by the expert has been identified as species $S_j$ by BEEWATCH users. The total number of times users have identified any submission as species $S_j$ is calculated by summing all the cells in that column ($C_j$). The total number of records in the database for species $S_i$ (as identified by the expert) are calculated by totaling all the cells in that row ($R_i$). Totaling all the cells in every row and column gives $N$, the number of user submitted identifications in the dataset.

These are the conditional odds $O$ for a hypothesis $H$, given independent evidence $E_1$ to $E_n$. The Hypothesis $H$ in this context is a possible species identity. Each evidence $E_i$ comes from a crowdsourced identification by a user of BEEWATCH. The odds depend on $O(H)$, the prior odds of the hypothesis $H$ (as not all species are equally abundant, a priori some are more likely than others before we have seen any user identifications) and $\Lambda$ terms, each of which updates the existing odds for $H$ based on the incoming evidence $E_i$, $E_2$, to $E_n$. Intuitively, the conditional odds for a hypothesis $H$ increase when the numerator of the $\Lambda$ term in (2), the likelihood of seeing this evidence $E_i$ for the hypothesis $H$, is high and the denominator, the likelihood of seeing this evidence $E_i$ for alternative hypotheses, is low.

We estimate the prior odds and the $\Lambda$ terms from a confusion matrix of species identifications of BEEWATCH users versus taxonomic experts. This is schematically presented in Fig. 2, where each cell $count(S_i, S_j)$ is the count of how frequently a photo-record identified as species $S_i$ by the expert (the hypothesis) has been identified as species $S_j$ by BEEWATCH users (the evidence). Fig. 5 (a) shows this confusion matrix with actual counts generated from the dataset. The expert identification is the correct hypothesis, for which user identifications provide evidence. The diagonal represents cases where the user identification matches the expert's, while off-diagonal cells represent cases where the user identification is providing evidence for a different species.

Referring again to Fig. 2, the prior probability $P$ for each species $S_i$ can be estimated as the relative abundance of the species in the records:

$$P(H = S_i) = \frac{R_i}{N} \tag{3}$$

The prior odds $O$ that a submission has species identity $S_i$ is by definition the ratio of the prior probability of the submission having identity $S_i$ to not having identity $S_i$:

$$O(H = S_i) = \frac{P(H = S_i)}{1 - P(H = S_i)} = \frac{R_i}{N - R_i} \tag{4}$$

With reference to Fig. 2, each $\Lambda$ term is estimated from the data as follows. The possibilities where the expert has identified a submission as $S_i$ are represented by the corresponding row total $R_i$. The conditional probability that an identification by a user is $S_j$ is then estimated as the proportion of this row that intersects with column $S_j$:

$$P(E = S_j | H = S_i) = \frac{count(S_i, S_j)}{R_i} \tag{5}$$

Similarly, the possibilities where the real species is not $S_i$ are represented by every row but the $i^{\text{th}}$ row; i.e. $N - R_i$. The conditional probability that an identification by a user is $S_j$ in this event is then estimated as the proportion of these rows that intersect with column $S_j$. Thus:

$$P(E = S_j|H \neq S_i) = \frac{C_j - count(S_i, S_j)}{N - R_i} \tag{6}$$

Finally, substituting (5–6) into the definition of a $\Lambda$ term (2), we obtain:

$$\Lambda(H = S_i|E = S_j) = \frac{count(S_i, S_j)}{R_i} \times \frac{N - R_i}{C_j - count(S_i, S_j)} \tag{7}$$

One such $\Lambda$ term is computed for each of the possible hypothesis/evidence pairs.

As each user identifies the photo independently to the others, the odds can be updated each time new evidence ($E_{n+1}$) comes in just by multiplying the existing odds (calculated from evidence $E_1$ to $E_n$) with the appropriate $\Lambda$ term. This follows directly from (1), and is derived in Appendix A:

$$O(H = S_j|E_1, ..., E_{n+1} = S_i) = O(H = S_j|E_1, ..., E_n) \times \Lambda(H = S_j|E_{n+1} = S_i) \tag{8}$$

The model we have described takes into account differences in the ease of identification of species by capturing the likelihoods of specific kinds of errors made by users, so that odds for each species can be updated based on any incoming user identification. However, the model as it stands averages over the behavior of all the users and fails to specifically model characteristics of individual users. We will refer to this as MODEL 1, The 'species model'. Instead of creating a single Fig. 2 aggregating data from all the users, we could instead create separate tables of counts from identifications by individual users. This would allow us to model the specific species identification errors made by individual users; i.e, take into account both the ease of identification of species and the abilities of different user. In practice, this requires the computation of separate $\Lambda_k(H = S_i|E = S_j)$ terms for each user $k$, by only considering identifications made by user $k$. We will refer to this as MODEL 2, the 'species+user model'. For MODEL 2, we compute the prior odds as before (as these are computed from species abundance and are user independent), but as each identification $S_j$ by a user $k$ come in, we multiply the odds for each hypothesis $H = S_i$ by the user specific $\Lambda_k(H = S_i|E = S_j)$, in contrast to MODEL 1 where we would have used the user-averaged $\Lambda(H = S_i|E = S_j)$.

### 3.4. Model smoothing

The critical issue for statistical models trained on a dataset is generalization: How accurate will the models' predictions be on previously unseen data? This is a particular concern for models with a large number of parameters, as these are difficult to estimate reliably from limited amounts of data. Consider again, (7), which is the calculation of a $\Lambda$ term. The numerator contains the term $count(S_i, S_j)$, the number of times a submission identified by the expert as $S_i$ is identified by users (a particular user for MODEL 2) as $S_j$. If this is zero then the lambda term $\Lambda(H = S_i, E = S_j)$ will be zero. Whenever this is used in the calculations for a submission to be identified (8), the odds for $S_i$ will become zero (and no matter what later user identifications are made, the odds for $S_i$ will remain at zero). In other words, if a user makes a mistake that has not been encountered in the dataset used for building the model, the correct identification can never be achieved. Similarly, the denominator in (7) contains the

term $C_j - count(S_i, S_j)$. If a species $S_i$ has only ever been identified by users as $S_j$ in the dataset (most likely, when $S_i = S_j$, i.e. a species has never been misidentified by a user, but also if $S_i \neq S_j$, if a species is always misidentified as the same other species), then $C_j - count(S_i, S_j) = 0$ and hence $\Lambda(H = S_i, E = S_j)$ becomes infinity, meaning that the consensus identification $S_i$ is now unbeatable. Now, it might be that two species are impossible to confuse, or that a particular species is impossible to misidentify, or indeed that a particular species is always misidentified the same way. However, it is more likely that certain 'expert/evidence combinations' have simply not been observed in the dataset and that both very low and very high $\Lambda$ values need to be moderated to assign small likelihoods to previously unseen events. This process is called smoothing.

We implemented Laplace smoothing, also called add-one smoothing [Simonoff 1995], by adding a count of $1$ in each cell in Fig. 2, and the corresponding tables for MODEL 2. This served to give a small probability to unseen evidence/hypothesis combinations and moderated the size of the $\Lambda$ terms; for example, preventing values of zero or infinity for $\Lambda$. For MODEL 2, generalization is an even more critical issue. As reported in Sec. 3.2, most users submitted fewer than ten identifications, but even regular users might only have covered a small subset of species, meaning that the unsmoothed lambda term $\Lambda_k(H = S_i|E = S_j)$ was zero for most combinations of $k$ (the user), $S_i$ and $S_j$. Thus, in addition to the smoothing described above, we further moderated MODEL 2 by combining it with the smoothed MODEL 1 in the ratio $3 : 1$; i.e. the smoothed values for MODEL 2 were $\frac{1}{4}\Lambda(H = S_i|E = S_j) + \frac{3}{4}\Lambda_k(H = S_i|E = S_j)$. We preferred this to the alternative community-based Bayesian label aggregation model [Venanzi et al. 2014], which creates confusion matrices for communities of similar users, because we had no evidence for the presence of such communities within BEEWATCH.

### 3.5. Model application

We applied both models by incrementally accepting crowdsourced identifications for a photo submission and updating the odds for each of the categories (hypotheses) until either (a) the odds for a category exceeded $9$ (a probability of $9/(9 + 1) = 0.90$ of the identification being correct; see Sec. 5.2 for a discussion of why this is an appropriate threshold to adopt), or (b) we ran out of user provided identifications, in which case the model failed to derive a consensus identification.

### 3.6. Model evaluation

The utility of the above-described consensus models in the biological recording arena was evaluated on the basis of three dimensions: (i) the size of the crowd needed to reach such consensus identifications (i.e. average number of identifications needed per photograph); (ii) the proportion of photographs for which the model succeeded in producing a consensus identification; and (iii) the proportion of consensus identifications made by the model that were correct, in total and on a by-species basis. We compared these models to two standard baselines: (1) a commonly used non-Bayesian consensus model, majority voting (MV), which accepts all the crowd identifications for a photo, treats these as votes, and selects the species which has collected the maximum number of votes from the crowd; and (2) the standard Multinomial Naive Bayes (MNB) classifier, as implemented in the Weka toolkit [Hall et al. 2009], which was found to outperform the default Naive Bayes classifier (based on a multi-variate Bernoulli event model) for this task.

Having set out the evaluation dimensions, we are interested in two different questions about the Bayesian models we described. First, how good are the models at classifying photo submissions as one of the $22$ species of bumblebees when it is known that such a classification is possible? Second, how good are the models when the dataset contains photo submissions that are too poor in quality to be identified to species level

or concern species outwith the target group (e.g. a hoverfly of solitary bee instead of a species of bumblebee), as is typically the case in citizen science initiatives? We shall consider the two cases separately in Sec. 4.1 and 4.2.

To evaluate our models, we really seek evidence about whether *in general* if they are 'trained' on one set of data (i.e. model parameters are estimated from this data) then they will perform well (when 'tested') on other data, which have not been used to build the models. To address this, we evaluated our models by performing ten-fold cross-validation, with the data repeatedly (i.e. $10\times$) partitioned into two subsets, a larger set being used for training (estimating the parameters in the model) and a smaller one for testing (evaluating the performance of the model). By repeating the partitioning multiple times and averaging results over the different partitionings, we reduced the variability of testing on small amounts of data in any single partitioning. To obtain a testing set in each 'fold', 10% of the photo submissions for which the full set of ten crowdsourced identifications had accumulated was sampled in such a manner that each of these submissions appeared in the test set in exactly one fold out of ten. In each fold, the remaining 90% of the photo-records with a full set of crowdsourced identification, as well as all the photo-records with fewer than 10 crowdsourced identifications, were then used to train the model.

## 4. RESULTS

### 4.1. Categorization as one of the 22 species of bumblebee

We filtered the dataset described in Sec. 3.2 to only contain photo submissions which the expert had positively identified as a bumblebee species and for which the full set of ten crowdsourced identifications had accumulated. This left $387$ records of $15$ bumblebee species (as identified by the expert) from a possible $22$.

Table I shows the results for MODEL 1 and MODEL 2 as well as the majority vote (MV) and Multinomial Naive Bayes (MNB) baselines. Across all species MODEL 1 succeeded in reaching consensus identification for 98% of photo submissions and MODEL 2 for $97\%$, comparable to MV ($97\%$). The MV algorithm failed to make a prediction in cases where no species managed a simple majority; i.e. when the top two species have the same number of votes. The Weka implementation of MNB by default makes predictions in all cases. For a fair comparison of accuracies in the table, we assigned a probability threshold for accepting a MNB prediction, such that consensus was achieved for the same proportion of records as MODEL 2; i.e. $97\%$.

It was striking how efficient the incremental Bayesian models were in reaching a consensus. On average, MODEL 1 needed only $2.6$ identifications, and for MODEL 2 this was $3.2$. Both the baselines (MV and MNB) used all the identifications available.

Both MODEL 1 and MODEL 2 fared well compared to the non-incremental baselines in terms of accuracy, despite requiring crowds that were a fraction of the size (measured by average number of identifications needed per photo-record). MODEL 2, which takes into account differences between both users and species, achieved the highest accuracy of $0.91$, a value in excess of the set threshold for accepting a consensus identification (an odds of $9$, which is equivalent to a probability of $9/(9+1) = 0.90$). The quality of the fit is further discussed in Sec. 5.1.

In the table, *recall* refers to the *detectability* of a species, whilst *precision* expresses how reliable a record of a certain species is. For example, MODEL 2 recall of *B. terrestris* was $0.91$ (Tab. I), meaning that this algorithm managed to detect $91\%$ of the records identified as that species by an expert (i.e. $9\%$ of *B. terrestris* records were mislabeled). MODEL 2 precision for this species was $0.76$, meaning that whenever the algorithm concluded *B. terrestris*, this was actually correct in $76\%$ of the cases (i.e. $24\%$ of records identified by the algorithm as *B. terrestris* were incorrect).

Table I. Results table for classification as one 22 bumblebee species in the UK, using 10-fold crossvalidation. For each bumblebee species (column 1, as IDed by the expert), Rec is the *recall* of that species, defined as the proportion of records of that species (as IDed by the expert) that are correctly identified by the model, and Prec is the *precision* for the species, defined as the proportion of records identified as that species by the model that are indeed that species (as IDed by the expert).

| | No. of | Majority Vote | | Naive Bayes | | MODEL 1 | | MODEL 2 | |
|---|---|---|---|---|---|---|---|---|---|
| Species | Records | Rec | Prec | Rec | Prec | Rec | Prec | Rec | Prec |
| *B. pascuorum* (common carder) | 106 | 0.95 | 0.99 | 0.99 | 0.99 | 0.99 | 0.97 | 0.99 | 0.98 |
| *B. pratorum* (early) | 67 | 0.78 | 0.98 | 0.85 | 0.95 | 0.86 | 0.93 | 0.90 | 0.95 |
| *B. terrestris* (buff-tailed) | 58 | 0.87 | 0.80 | 0.77 | 0.80 | 0.86 | 0.75 | 0.91 | 0.76 |
| *B. hypnorum* (tree) | 56 | 0.98 | 0.96 | 0.98 | 0.98 | 0.96 | 0.98 | 0.96 | 0.98 |
| *B. lapidarius* (red-tailed) | 36 | 0.92 | 0.97 | 0.94 | 0.92 | 0.94 | 0.86 | 0.97 | 0.92 |
| *B. lucorum* (white-tailed) | 26 | 0.75 | 0.62 | 0.81 | 0.58 | 0.65 | 0.65 | 0.65 | 0.71 |
| *B. hortorum* (garden) | 23 | 0.78 | 1.00 | 0.95 | 0.91 | 0.87 | 0.91 | 0.83 | 0.95 |
| *B. monticola* (bilberry) | 3 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| *B. vestalis* (southern cuckoo) | 2 | 0.50 | 1.00 | 0.00 | - | 0.50 | 0.50 | 0.50 | 1.00 |
| *B. jonellus* (heath) | 2 | 0.00 | - | 0.00 | - | 0.00 | - | 0.00 | - |
| *B. bohemicus* (gypsy cuckoo) | 2 | 1.00 | 0.67 | 1.00 | 0.33 | 0.50 | 0.33 | 1.00 | 0.67 |
| *B. rupestris* (red-tailed cuckoo) | 2 | 1.00 | 1.00 | 0.00 | - | 0.00 | - | 0.00 | - |
| *B. ruderarius* (red-shanked carder) | 2 | 0.50 | 1.00 | 0.00 | - | 0.00 | - | 0.50 | 1.00 |
| *B. distinguendus* (great yellow) | 1 | 0.00 | - | 0.00 | - | 0.00 | - | 0.00 | - |
| *B. sylvestris* (forest cuckoo) | 1 | 1.00 | 1.00 | 0.00 | - | 0.00 | - | 0.00 | - |
| Average no. of IDs needed per photo | | 10.66 | | 10.66 | | 2.58 | | 3.22 | |
| Average accuracy over dataset | | 0.88 | | 0.90 | | 0.88 | | 0.91 | |
| Proportion for which consensus achieved | | 0.97 | | 0.97 | | 0.98 | | 0.97 | |

Both our models also proved robust to the order in which individual identifications were processed. When the order of processing identifications was randomized $10$ times, the average number of identifications needed per photo for MODEL 1 ranged between $2.58$ and $2.76$, and the average accuracy between $0.87$ and $0.90$. For MODEL 2, the respective ranges were $3.20 - 3.43$ and $0.89 - 0.92$. The proportion of photos attempted ranged between $0.98$ and $0.99$ for MODEL 1 and was consistently $0.97$ for MODEL 2.

### 4.2. Ability of the models to filter out unusable records

As described earlier, one peculiarity of photo-based citizen science initiatives is that submitted photos may not contain the target species group or if they do, images may be too poor in quality to be identifiable to species level. Indeed, in our case a considerable number of photo submissions ($13.6\%$) did not concern bumblebees (but often hoverflies, flower flies and solitary bees), and a further $21.2\%$ of photo-submissions were of insufficient quality for reliable identification of the species of bumblebee, even by taxonomic experts. Our dataset therefore also contains records labeled with one of two additional categories, namely 'not a bumblebee' and 'not identifiable'. If such records cannot be filtered out before submission to the consensus module, then the consensus module would need to handle such records too.

Table II shows the performance of the four models for classifying images as either one of the $22$ focal species, or as one of 'not a bumblebee' or 'not identifiable', using the dataset of $594$ photo-submissions containing at least 10 crowd identifications as described in Sec. 3.2. MODEL 1 needed on average only $3.9$ identifications to arrive at a consensus, while MODEL 2 required $4.3$ identifications. The MV and MNB baselines again used all the identifications available.

MODEL 2, which took into account differences between species and users achieved the highest average accuracy of $0.80$, though it arrived at consensus for fewer photos than MV. As before, for a fair comparison of accuracies, we set a probability threshold for accepting a MNB prediction in a manner that consensus was achieved for the same proportion of records as MODEL 2; i.e. $94\%$. (For a fairer comparison of the models, we will later discuss a) the trade-off between accuracy and the proportion for which

Table II. Results table for classification as 'not a bumblebee', 'not identifiable' or one of the 22 species, using 10-fold crossvalidation. For each bumblebee species (column 1, as IDed by the expert), Rec is the *recall* of that species, defined as the proportion of records of that species (as IDed by the expert) that are correctly identified by the model, and Prec is the *precision* for the species, defined as the proportion of records identified as that species by the model that are indeed that species (as IDed by the expert).

| | No. of | Majority Vote | | Naive Bayes | | MODEL 1 | | MODEL 2 | |
|---|---|---|---|---|---|---|---|---|---|
| Species | Records | Rec | Prec | Rec | Prec | Rec | Prec | Rec | Prec |
| *B. pascuorum* (common carder) | 106 | 0.95 | 0.86 | 0.98 | 0.86 | 0.99 | 0.83 | 0.98 | 0.84 |
| *B. pratorum* (early) | 67 | 0.78 | 0.93 | 0.87 | 0.89 | 0.86 | 0.89 | 0.89 | 0.89 |
| *B. terrestris* (buff-tailed) | 58 | 0.87 | 0.61 | 0.76 | 0.67 | 0.79 | 0.58 | 0.88 | 0.64 |
| *B. hypnorum* (tree) | 56 | 0.98 | 0.90 | 0.96 | 0.95 | 0.96 | 0.93 | 0.96 | 0.93 |
| *B. lapidarius* (red-tailed) | 36 | 0.92 | 0.89 | 0.94 | 0.85 | 0.97 | 0.82 | 0.97 | 0.85 |
| *B. lucorum* (white-tailed) | 26 | 0.75 | 0.44 | 0.77 | 0.48 | 0.65 | 0.44 | 0.69 | 0.62 |
| *B. hortorum* (garden) | 23 | 0.78 | 0.55 | 0.90 | 0.54 | 0.91 | 0.54 | 0.86 | 0.61 |
| *B. monticola* (bilberry) | 3 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| *B. jonellus* (heath) | 2 | 0.00 | - | 0.00 | - | 0.00 | - | 0.00 | - |
| *B. bohemicus* (gypsy cuckoo) | 2 | 1.00 | 0.50 | 1.00 | 0.20 | 1.00 | 0.50 | 0.50 | 1.00 |
| *B. vestalis* (southern cuckoo) | 2 | 0.50 | 0.20 | 0.00 | - | 0.50 | 0.20 | 1.00 | 0.20 |
| *B. rupestris* (red-tailed cuckoo) | 2 | 1.00 | 1.00 | 0.00 | - | 0.00 | - | 0.00 | - |
| *B. ruderarius* (red-shanked carder) | 2 | 0.50 | 1.00 | 0.00 | - | 0.00 | - | 0.50 | 1.00 |
| *B. sylvestris* (forest cuckoo) | 1 | 1.00 | 0.50 | 0.00 | - | 1.00 | 1.00 | 0.00 | - |
| *B. distinguendus* (great yellow) | 1 | 0.00 | - | 0.00 | - | 0.00 | - | 0.00 | - |
| Not a bumblebee | 81 | 0.85 | 0.91 | 0.92 | 0.88 | 0.87 | 0.87 | 0.89 | 0.88 |
| Not identifiable | 126 | 0.34 | 0.67 | 0.32 | 0.75 | 0.26 | 0.90 | 0.39 | 0.83 |
| Average no. of IDs needed per photo | | 10.57 | | 10.57 | | 3.89 | | 4.31 | |
| Average accuracy over dataset | | 0.76 | | 0.78 | | 0.76 | | 0.80 | |
| Proportion for which consensus achieved | | 0.97 | | 0.94 | | 0.94 | | 0.94 | |

consensus is reached, and b) how dependent performance is on the identification skills of individual participants).

In line with this being a more difficult (but realistic) task, all four models presented lower values for individual species (Tab. II) than we were seeing previously (Tab. I), and as a consequence, the average accuracy for each model was also lower than before. This indicates, unsurprisingly, that the submission of photo material which is of insufficient quality considerably hampers crowdsourced identification. The reasons for this drop in accuracy (for MODEL 2, from $91\%$ for the $22$ category case to $80\%$ for the $24$ category case) are discussed next.

## 5. DISCUSSION OF RESULTS

### 5.1. Quality of parameter fitting within the models

Underfitting, which results from insufficient data to obtain good estimates for all the model parameters, and overfitting, where the model learns statistical patterns with no validity (often called noise) from the data as well as valid patterns, are well studied in statistical modeling. The consequence of both is that performance of the model on unseen data is lower than it would be on the data used to build the models.

In the first evaluation, which assumed a clean dataset where it is possible to identify each photo submissions as one of the 22 species of bumblebee, we do not see evidence of over or under fitting. The accuracy of MODEL 2 is $91\%$ on unseen data, comparable to the $90\%$ likelihood threshold we used to accept a record. This suggests that our model (the Bayesian reasoning together with the smoothing procedures) is estimating the model parameters (prior odds and lambda terms) accurately.

However, the accuracies we observe for the second statement of the problem, where the dataset additionally contains photo-submissions that are not identifiable even by the experts, are lower than expected ($80\%$ for MODEL 2). To explore this further, Fig. 5 (b) shows the confusion matrix between expert identifications and those made by MODEL 2 with odds threshold of $9$. As can be seen, much of the noise in individ-

ual identifications (Fig. 5 (a)) is eliminated, leaving few cases of misclassification (i.e. away from the diagonal) other than those concerning confusion between a bumblebee species and the 'not identifiable' category. Confusions between two species or a species and 'not a bumblebee' occur rarely ($5.0\%$ of records).

As is clear from Fig. 5 (b), the drop in accuracy of the model is mostly due to mislabeling one of the $22$ species as 'not identifiable', or vice versa. Intuitively, users might misidentify one bumblebee species as another bumblebee species because there are visual similarities between the species. Such patterns of misclassification when learned can generalize to new photographs of the species involved. One reason why our parameter estimates involving the 'not identifiable' are poor is that the categorization of a photo as 'not identifiable' has to do with the quality of the photograph, not on the visual similarity of some prototypical 'not identifiable' category to a bumblebee species. Thus the patterns learned by the models for misclassifications between a species and the 'not identifiable' category do not generalize well to new data. For these reasons, we have to accept that for datasets with potentially unidentifiable records, the models deliver accuracies below those expected from the odds thresholds we set.

### 5.2. Quality assurance

The nature of crowdsourcing is that there is unfortunately no consensus model guaranteeing perfect identification. It could be argued that scientific recording cannot tolerate any errors, but on the other hand, because human error is inevitable in any system, it is clear that no existing recording program can guarantee this. Indeed, even taxonomic experts do not always agree on the correct classification of a species in a photo-record. We invited two bumblebee experts from the Bumblebee Conservation Trust to independently identify 47 randomly selected photo-records using our interface. Their accuracies on the task were $85.1\%$ and $87.2\%$ (when evaluated against the official BEEWATCH record validated by a taxonomic expert for the program). That the identification accuracy of even experts in that trial was not 100% illustrates the difficulty of identification from photographs alone, with no means of investigating ID features out of shot or too small to see clearly in the photographs provided. We will discuss the implications for biological recording later, but first we will discuss the performance of the models when their parameters are adjusted to achieve differing levels of accuracy.

*5.2.1. Increasing the confidence in consensus identification.* All four models contain parameters that can be adjusted to increase or decrease the confidence in the consensus identification. For the Bayesian models, we can adjust the odds or probability threshold at which a consensus identification is accepted. For the majority vote baseline, we can specify a threshold for the 'margin of victory'; i.e. the difference in the number of votes between the majority identification and its nearest competitor. Figure 3 shows how increasing the threshold for consensus results in higher accuracy but smaller proportion of photos for which consensus is reached. The performance of MODEL 2 is comparable to Multinomial Naive Bayes (MNB) and performs consistently better than the other two; for any accuracy level specified, MODEL 2 and MNB achieve consensus for a greater proportion of photos than the other two models. For instance, if required to have an accuracy of $90\%$, MODEL 2 and MNB can identify $61\%$ of records compared to $51\%$ for MV, a relative increase of $20\%$.

*5.2.2. Handling crowds of different quality.* Since the Bayesian models take into account specific types of errors, we would expect their power to be greater (relative to MV) as the quality of the crowd gets worse. Indeed it has been reported previously for tasks with 2–4 labels [Sheshadri and Lease 2013] that MV shows little resilience to even modest noise levels of 25% and is thus unsuited for many consensus tasks.

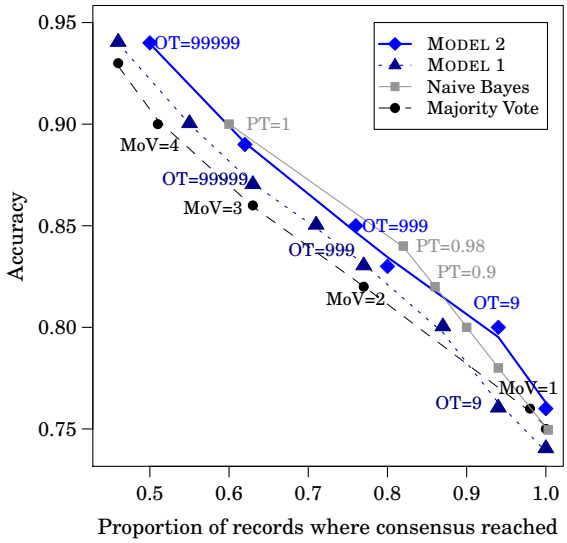



Fig. 3. Plots showing the tradeoff between accuracy and proportion for which consensus is reached, as the threshold for consensus (odds threshold (OT) for Incremental Bayesian models, probability threshold (PT) for MNB, and margin-of-victory (MoV) for majority vote) is increased. The right-most point for each model was obtained by labeling all photos for which no consensus was achieved with the label 'not identifiable', which resulted in each model making a prediction for every photo. MODEL 2 and MNB consistently outperform the other two; for any accuracy level specified, they achieve consensus for a greater proportion of photos than the other two models.

Figure 4 plots the proportion of photos for which consensus is achieved, when the threshold for consensus (odds/probability thresholds for Bayesian models and margin of victory for majority voting) is adjusted to achieve two different quality guarantees (80% and 90% accuracy), as a function of crowd quality. In this plot the accuracy of individual labels is varied by introducing errors at random to the left of the vertical dotted line, which marks the average label accuracy for our dataset, and correcting errors at random to the right of the vertical dotted line. As expected, when the accuracy of individual labels reduces, the Bayesian models outperform majority vote by increasing margins. In contrast, as individual labels get more accurate, the models converge. When individual label accuracy is greater than 55%, there is little to separate MODEL 2 from MNB; however MNB is more robust to deteriorating crowd quality at the point where individual label accuracy drops to 10% points below that observed in BEEWATCH. These graphs confirm that the incremental Bayesian Model 2 provides similar performance to MNB for the individual label accuracies we expect in biological recording, even while minimizing crowd size. They also confirm that MODEL 2 outperforms the commonly used majority vote model, which we conclude like [Sheshadri and Lease 2013] is not well suited to consensus tasks with moderate to heavy noise.

*5.2.3. Error reduction through consensus.* Figure 5 (a) shows how users have identified photo-records compared to the expert, showing a considerable number of misidentifications by users (i.e. counts off the diagonal). Figure 5 (b) illustrates how our Bayesian MODEL 2 (with default odds threshold of 9) eliminates much of the noise in Fig. 5 (a), leaving few cases of misclassification (i.e. away from the diagonal) other than those concerning confusion between a bumblebee species and the 'not identifiable' category.

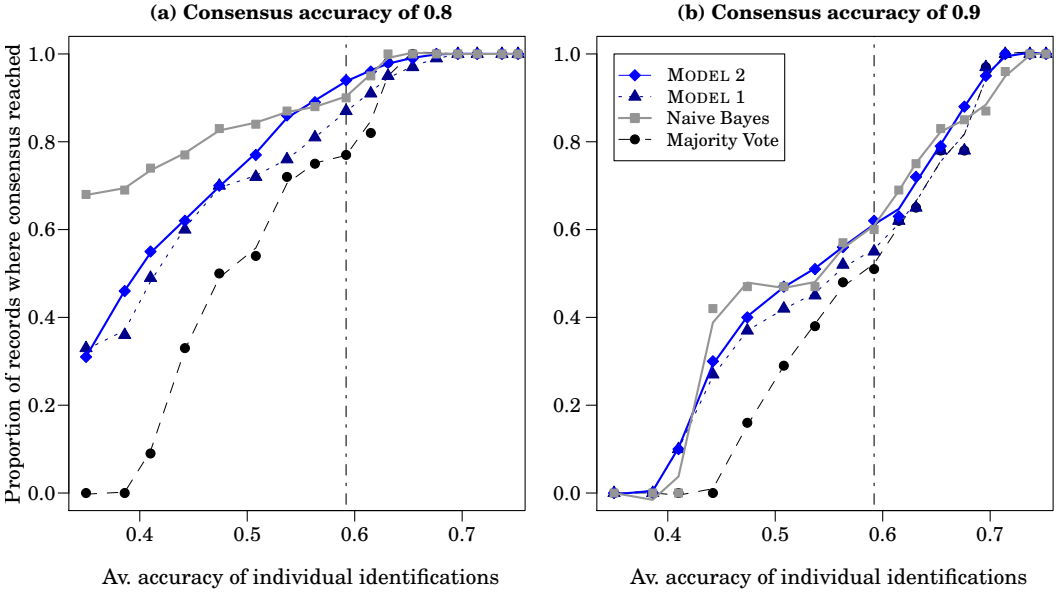



Fig. 4. Plots showing how the performance of the models deteriorates with lower identification accuracies by individual users. The X axis varies the accuracy of individual identifications in the data set by randomly introducing or correcting (to the left or right respectively of the vertical dotted line that represents the individual label accuracy in our data set) 5%, 10%, 15%, 20%, 25%, 30%, 35% and 40% errors in individual labels. The Y axis shows the proportion of records for which consensus is achieved, when the threshold for consensus (OT, PT or MoV) has been adjusted to give an accuracy of (a) 80% and (b) 90% for the dataset.

The right-most column (excluding the bottom most cell) depicts instances where the model fails to make a species identification, when it was possible for the expert to do so (1.6% of records). The bottom row (excluding the right most cell) depicts instances where the model makes a species identification even though the expert has decided the specimen is not identifiable (11.3% of records). Confusions between two species or a species and 'not a bumblebee' occur rarely (5.0% of records). Fig. 5 (c) shows how further error reduction can be achieved by increasing the odds threshold to 9999. Now, confusions between two species occur for only 1.6% of records, with the remaining errors (8.3% of records) involving confusion between a bumblebee species and the 'not identifiable' category.

### 5.3. The need for consensus models in biological recording

Environmental concern, enforced by international policy obligations, has raised the demand for species distribution (and abundance) data well beyond the capacity of professional biologists to deliver this [Danielsen et al. 2005]. Hence, the general public are increasingly encouraged to act as biological recorders, particularly for species groups such as pollinating insects that are important to society, but for which few expert recorders exist.

The growth in camera ownership (e.g. in mobile phones) has meant citizen science initiatives increasingly rely on citizen recorders submitting photos, which are then verified by relatively few experts able to reliably identify the species. If popular, such initiatives are put at immediate risk of collapsing under the large number of submissions due to a lack of available expert time, especially if short response times are required to keep recorders engaged. Numerous biological recording schemes are ad-

| Expert \ | Users | *B. terrestris* | *B. pratorum* | *B. pascuorum* | *B. lucorum* | *B. lapidarius* | *B. hortorum* | *B. jonellus* | *B. hypnorum* | *B. monticola* | *B. sylvestris* | *B. campestris* | *B. soroeensis* | *B. bohemicus* | *B. vestalis* | *B. barbutellus* | *B. rupestris* | *B. humilis* | *B. muscorum* | *B. ruderatus* | *B. ruderarius* | *B. sylvarum* | *B. distinguendus* | Not a bumblebee | Not identifiable |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *B. terrestris* | 364 | 35 | | 63 | 3 | 11 | | 1 | | 1 | 4 | 8 | 2 | | | 3 | | | | | | | 15 | 59 |
| *B. pratorum* | 76 | 399 | 2 | 37 | 47 | 3 | | 3 | 10 | 6 | 13 | 5 | 1 | 2 | | 3 | | | 3 | 2 | | 1 | 13 | 87 |
| *B. pascuorum* | 6 | 11 | 718 | 1 | | | | 73 | | | 5 | | 1 | | | 2 | 26 | 33 | 1 | | 1 | | 30 | 90 |
| *B. lucorum* | 67 | 4 | | 151 | | 5 | 2 | 4 | | | | 6 | 6 | | | | | | | | | 1 | | 18 |
| *B. lapidarius* | 3 | 20 | 2 | 2 | 253 | 1 | | | | 1 | 6 | | | | | 21 | | | | 26 | | | 2 | 37 |
| *B. hortorum* | 22 | 2 | 1 | 52 | | 138 | 16 | | | | 2 | | 4 | | | | | | 6 | | | | | 15 |
| *B. jonellus* | | | | | | | | | | | | | | | | | | | | | | | | |
| *B. hypnorum* | | | 13 | | | | 1 | 536 | | | 1 | | | | 2 | | | | 6 | | | | 3 | 26 |
| *B. monticola* | | | | | | | | | 16 | | | | | | | 4 | | | | | | | 1 | 1 |
| *B. sylvestris* | | | | 2 | | | | | | 6 | | | | 1 | | | | | | | | | | 1 |
| *B. campestris* | | | | | | | | | | | | | | | | | | | | | | | | |
| *B. soroeensis* | | | | | | | | | | | | | | | | | | | | | | | | |
| *B. bohemicus* | | | | | | | | | | | | | | | | | | | | | | | | |
| *B. vestalis* | | | | | | | | | | | | | | | | | | | | | | | | |
| *B. barbutellus* | | | | | | | | | | | | | | | | | | | | | | | | |
| *B. rupestris* | | | | | 10 | | | | | | | | | | | 12 | | | | | | | | |
| *B. humilis* | | | | | | | | | | | | | | | | | | | | | | | | |
| *B. muscorum* | | | | | | | | | | | | | | | | | | | | | | | | |
| *B. ruderatus* | | | | | | | | | | | | | | | | | | | | | | | | |
| *B. ruderarius* | | | | | | | | | | | | | | | | | | | | | | | | |
| *B. sylvarum* | | | | | | | | | | | | | | | | | | | | | | | | |
| *B. distinguendus* | | | | 3 | | | | | | | | | | | | | 1 | | | | 2 | 3 | | 2 |
| Not a bumblebee | 1 | 2 | 70 | 2 | 20 | | | 13 | 10 | | 22 | 1 | | 3 | 5 | 6 | 5 | 18 | 7 | 5 | 2 | 494 | 128 | |
| Not identifiable | 234 | 48 | 97 | 176 | 6 | 141 | 21 | 20 | | 14 | 17 | 5 | 21 | 18 | 7 | 1 | 5 | 9 | 29 | | 1 | | 65 | 345 |

(a) Expert vs User identifications.

| Expert \ | MODEL 2 | *B. terrestris* | *B. pratorum* | *B. pascuorum* | *B. lucorum* | *B. lapidarius* | *B. hortorum* | *B. hypnorum* | *B. monticola* | *B. bohemicus* | *B. vestalis* | *B. rupestris* | *B. ruderarius* | Not a bumblebee | Not identifiable |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *B. terrestris* | 50 | 2 | | 2 | | | | | | | | | | 1 | 1 |
| *B. pratorum* | 1 | 53 | | | 1 | | | | | | | | | 1 | 4 |
| *B. pascuorum* | | | 100 | | | | 1 | | | | | | | | 1 |
| *B. lucorum* | 8 | | | 17 | | | | | | | | | | | |
| *B. lapidarius* | | 1 | | | 33 | | | | | | | | | | |
| *B. hortorum* | | | | 1 | | 19 | | | | | | | | | 1 |
| *B. hypnorum* | | | 1 | | | | 53 | | | | | | | | |
| *B. monticola* | | | | | | | | 3 | | | | | | | |
| *B. bohemicus* | | | | | | | | | 1 | | | | | | 1 |
| *B. vestalis* | | | | | | | | | | 1 | | | | | |
| *B. rupestris* | | | | | 1 | | | | | | 1 | | | | |
| *B. ruderarius* | | | | | 1 | | | | | | | 1 | | | |
| Not a bumblebee | | | 4 | | 1 | | 1 | | | | | | 68 | 1 | |
| Not identifiable | 17 | 3 | 13 | 5 | 1 | 12 | 2 | | | 4 | | | 6 | 37 | |

(b) Expert vs MODEL 2, $OT = 9$.

| Expert \ | MODEL 2 | *B. terrestris* | *B. pratorum* | *B. pascuorum* | *B. lucorum* | *B. lapidarius* | *B. hortorum* | *B. hypnorum* | *B. monticola* | *B. vestalis* | *B. rupestris* | *B. ruderarius* | Not a bumblebee | Not identifiable |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *B. terrestris* | 27 | | | | | | | | | | | | | |
| *B. pratorum* | 1 | 36 | | | | | | | | | | | | |
| *B. pascuorum* | | | 86 | | | | 1 | | | | | | | |
| *B. lucorum* | 1 | | | 5 | | | | | | | | | | |
| *B. lapidarius* | | 1 | | | 31 | | | | | | | | | |
| *B. hortorum* | | | | | | 11 | | | | | | | | |
| *B. hypnorum* | | | | | | | 51 | | | | | | | |
| *B. monticola* | | | | | | | | 3 | | | | | | |
| *B. vestalis* | | | | | | | | | 1 | | | | | |
| *B. rupestris* | | | | | 1 | | | | | | | | | |
| *B. ruderarius* | | | | | 1 | | | | | | | | | |
| Not a bumblebee | | | | | | | | | | | | | 54 | |
| Not identifiable | 6 | 1 | 8 | 2 | | 3 | 2 | | 3 | | | | 3 | 4 |

(c) Expert vs MODEL 2, $OT = 9999$.

Fig. 5. Confusion Matrices. (a) The accuracy of individual identifications is 59%, resulting in a wide distribution of errors (off-diagonal cells). (b) For an odds threshold of $OT = 9$, the accuracy of MODEL 2 is 80%, resulting in visible error correction (fewer counts in off-diagonal cells), but consensus is achieved for only 94% of photos. (c) For an odds threshold of $OT = 9999$, MODEL 2 shows greater capacity to reduce the errors made by individual users (89% accuracy), but achieves consensus for a smaller proportion of photos (62%).

ministered by very few staff, notably when dealing with charismatic species groups. As an example, as many as 1613 photos were submitted to BEEWATCH over a single 20 day period, with only two experts at hand that were employed part-time to handle photo submissions.

There is a growing demand for citizen-science data which can be used in scientific research [Danielsen et al. 2005; Comont et al. 2012; Roy et al. 2012] and to influence policy (http://ec.europa.eu/environment/integration/research/newsalert/pdf/IR9.pdf), which means that scale, timeliness and accuracy are all important. Many of the techniques used to date to alleviate the problem of a lack of expert time have traded one of these in favor of one of the others; for example, Friends of the Earth's Great British Bee Count (http://greatbritishbeecount.co.uk/) accumulated 820,000 sightings in three months, but with no attempt at species level identification or verification, thus severely limiting the use of such data for scientific purposes.

Inviting lay people to not only submit photographs but also take part in the identification process can be a way to reduce the burden on the few experts, freeing them to prioritize rare species (for which crowd sourcing with random sampling does not provide many records) or difficult photographs (where the crowd does not agree). This will allow the initiative to flourish, if procedures are in place to swiftly and effectively derive at consensus identification of photographed specimens. Consensus models are in this sense a new practical tool to be added to the armory of techniques having to be adopted by current recording programs. However, consensus models require the recruitment of a 'crowd' of people able to make relevant identification decisions. A crowdsourcing program will itself be limited by the size of crowd required and the quality and volume of work that the crowd members can offer.

The Bayesian models developed here address this situation head-on. As a new identification by a user comes in, a decision is made on whether to accept the consensus identification or solicit further identifications by users, based on the accuracy required. The power of our models is threefold. First, they account for species-specific differences in the ease of identification (MODELS 1 & 2) as well as differential skill level among users (MODEL 2). Second, the existence of common identification mistakes allows a crowdsourced identification for one species to provide varying levels of evidence for one or more other species. Thus, mistakes by users are effectively harnessed by the models to arrive at consensus: an important attribute for programs that rely on members of the public with varying identification skills. Third, the models re-evaluate the consensus identification with every new identification coming in; this minimizes the number of users to be consulted. Both incremental Bayesian models, but particularly MODEL 2, outperform a traditional majority vote approach, both in terms of the quality of the results and the size of the crowd required.

Reducing crowd size and thus our Bayesian approach is of utmost importance to the sustainability of photo-submission based recording schemes. For BEEWATCH to run without experts, every time a participant submits a new photo, they would, with MODEL 2, also need to help identify bumblebees for 3–4 photos submitted by others, in addition to their own submission, for there to be a balance between photo-submissions and consensus labeling. As we reported in Sec. 3.2, BEEWATCH averaged $5.48$ identifications per photo. This indicates the great importance of models such as ours that make efficient use of the scarce resource that is the 'crowdsourcer'.

### 5.4. Handling imperfect data within a biological recording scheme

In practice, a biological recording scheme has to accept that there will be *some* errors in the data it collects – it is simply a matter of what level of error is acceptable. This has to be balanced against the level of resources available and issues such as the value of the data for informing policy (in our case, for example, the UK's National Pollinator Strategy). In many situations, it may be better to have more data, or even any data at all, with a quantified level of inaccuracy, rather than to insist on unreasonably high levels of accuracy and have no data to work with.

It is worth discussing the quality of data generated by the consensus model in this context. Our results indicate that Bayesian consensus models can produce data that is as reliable as identification by experts, for the vast majority of submissions, without expert involvement ($85\%$ accuracy for identifying $72\%$ of submissions; see Fig. 3). It is possible with our models to specify a desired level of accuracy over a dataset. Figure 4 showed how this affects the proportion of photos that can be identified by consensus as a function of crowd quality. Bayesian models outperform majority voting by a wider margin when the quality of the crowd is worse. They are therefore particularly suited to initiatives that solicit records of species groups where knowledge among members of the general public is relatively low, or where species are relatively difficult to identify.

While our crowd sourcing methods derive highly likely identifications in the overwhelming number of cases, recording schemes are likely to want to apply more stringent rules for certain species, such as rarities or particularly difficult species. Crowdsourcing can also help identify these priority records; for example, all photos where the crowd consensus was a rare species, or even those where at least one of the crowd had identified the specimen as a rare or difficult-to-identify species could be prioritized for cross-checking by dedicated experts, easing the strain by essentially removing easy-to-identify, common species from the expert's workload.

## 6. CONCLUSIONS

We have unfolded an incremental Bayesian method for obtaining consensus on crowdsourcing tasks which takes into account both individual users' identification skills and the level of difficulty associated with identifying individual species. The model is suitable for classification tasks with many categories and also for relatively difficult tasks where participants exhibit a variety of skill levels. When testing our model on 'real-world' citizen science data, it outperformed the traditional 'majority vote' approach both in terms of the accuracy of obtained consensus identifications and the size of the crowd required. It also achieved similar accuracy to the non-incremental Naive Bayes model despite requiring smaller crowd sizes. Our findings demonstrate that a relatively complex identification task can be performed reliably through employing Bayesian methods with very small crowd sizes of around 3–5 users; i.e. through group sourcing. As the data collection exercise reported in Sec. 3.2 achieves on average $5.48$ crowdsourced identifications per submission, we can conclude that Bayesian methods are well suited for such applications, and indeed that such efficiency is *required*.

Crowdsourcing has to be planned within a wider environment in which the 'crowd' participates. In BEEWATCH, although we are asking members of the public to carry out a non-trivial classification task, we are providing support for that activity, in terms of an interactive online identification guide and automatically generated feedback on their identifications. We also provide an online training tool where users can practice, and links to further online resources. All of this helps users improve their identification skills, which further motivates them to participate.

It should be recognized that for much of science the dichotomy between 'expert' and 'lay' is essentially false. For biological recording, identification skills are on a continuum and schemes such as BEEWATCH by definition attract those members of the general public interested in the subject matter. In particular, there is considerable scope for explicit training of recorders (e.g. the training tool) as well as passive learning through feedback and practice. The advantage of the models presented here is that they can take this spectrum of identification ability into account when constructing a consensus ID, as well as the per-species level of identification difficulty. All users can contribute towards building consensus on a species, irrespective of the quality of their identification skills. As users get more experienced in identification, the performance of the consensus model can be expected to improve (see Fig. 4).

To conclude, crowdsourcing does not have to be used as an 'all or nothing' approach. Consensus models such as those presented here allow for the setting of parameters that balance the accuracy of the results against the size of crowd required, thus providing a system for the triage of records to prioritize the use of limited quantities of expert time. For biological recording tasks, Bayesian models estimate bias accurately because mistakes are not random: participants are typically motivated and participate because they care about nature and want to increase their knowledge. Therefore mistakes involve genuinely confusable species and can thus be modeled, or poor photo quality, resulting in confusion with the 'not identifiable' label. In other domains, there can be the risk of users acting maliciously, who do not attempt the task with the intention of helping – further research is needed to explore whether incremental models can be robust to such users.

**Acknowledgment**

This research is supported by an award made by the RCUK Digital Economy program to the University of Aberdeen's dot.rural Digital Economy Hub (ref. EP/G066051/1).

## REFERENCES

Steven Blake, Advaith Siddharthan, Hien Nguyen, Nirwan Sharma, Anne-Marie Robinson, Elaine O'Mahony, Ben Darvill, Chris Mellish, and René van der Wal. 2012. Natural Language Generation for Nature Conservation: Automating Feedback to help Volunteers identify Bumblebee Species. In *24th International Conference on Computational Linguistics (COLING 2012)*. Mumbai, India, 311–324.

David N Bonter and Caren B Cooper. 2012. Data validation in citizen science: a case study from Project FeederWatch. *Frontiers in Ecology and the Environment* 10, 6 (2012), 305–307.

Doug Clow and Elpida Makriyannis. 2011. iSpot Analysed: Participatory Learning and Reputation. In *Proc. the 1st International Conference on Learning Analytics and Knowledge (LAK '11)*. ACM, 34–43.

Richard F. Comont, Helen E. Roy, Owen T. Lewis, Richard Harrington, Christopher R. Shortall, and Bethan V. Purse. 2012. Using biological traits to explain ladybird distribution patterns. *Journal of Biogeography* 39, 10 (2012), 1772–1781. DOI:http://dx.doi.org/10.1111/j.1365-2699.2012.02734.x

Finn Danielsen, Neil D Burgess, and Andrew Balmford. 2005. Monitoring matters: examining the potential of locally-based approaches. *Biodiversity & Conservation* 14, 11 (2005), 2507–2542.

Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied statistics* 28, 1 (1979), 20–28.

Pinar Donmez and Jaime G Carbonell. 2008. Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In *Proc. the 17th ACM conference on Information and knowledge management*. ACM, 619–628.

Jeremy JD Greenwood. 2007. Citizens, science and bird conservation. *Journal of Ornithology* 148, 1 (2007), 77–124.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter* 11, 1 (2009), 10–18.

Wesley M Hochachka, Daniel Fink, Rebecca A Hutchinson, Daniel Sheldon, Weng-Keen Wong, and Steve Kelling. 2012. Data-intensive science applied to broad-scale citizen science. *Trends in ecology & evolution* 27, 2 (2012), 130–137.

Panagiotis G Ipeirotis, Foster Provost, Victor S Sheng, and Jing Wang. 2014. Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery* 28, 2 (2014), 402–441.

Ece Kamar, Severin Hacker, and Eric Horvitz. 2012. Combining human and machine intelligence in large-scale crowdsourcing. In *Proc. the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 467–474.

David R Karger, Sewoong Oh, and Devavrat Shah. 2014. Budget-optimal task allocation for reliable crowdsourcing systems. *Operations Research* 62, 1 (2014), 1–24.

Hongwei Li, Bo Zhao, and Ariel Fuxman. 2014. The wisdom of minority: discovering and targeting the right group of workers for crowdsourcing. In *Proc. the 23rd international conference on World wide web*. International World Wide Web Conferences Steering Committee, 165–176.

Chris J Lintott, Kevin Schawinski, Anže Slosar, Kate Land, Steven Bamford, Daniel Thomas, M Jordan Raddick, Robert C Nichol, Alex Szalay, Dan Andreescu, and others. 2008. Galaxy Zoo: morphologies

derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society* 389, 3 (2008), 1179–1189.

Greg Little, Lydia B Chilton, Max Goldman, and Robert C Miller. 2009. Turkit: tools for iterative tasks on mechanical turk. In *Proc. the ACM SIGKDD workshop on human computation*. ACM, 29–30.

Babak Loni, Jonathon Hare, Mihai Georgescu, Michael Riegler, Xiaofei Zhu, Mohamed Morchid, Richard Dufour, and Martha Larson. 2014. Getting by with a little help from the crowd: Practical approaches to social image labeling. In *Proc. the 2014 International ACM Workshop on Crowdsourcing for Multimedia*. ACM, 69–74.

Andrew Mao, Ece Kamar, and Eric Horvitz. 2013. Why Stop Now? Predicting Worker Engagement in Online Crowdsourcing. In *First AAAI Conference on Human Computation and Crowdsourcing*.

Helen E. Roy, Tim Adriaens, Nick J. B. Isaac, Marc Kenis, Thierry Onkelinx, Gilles San Martin, Peter M. J. Brown, Louis Hautier, Remy Poland, David B. Roy, Richard Comont, Ren Eschen, Robert Frost, Renate Zindel, Johan Van Vlaenderen, Oldich Nedvd, Hans Peter Ravn, Jean-Claude Grgoire, Jean-Christophe de Biseau, and Dirk Maes. 2012. Invasive alien predator causes rapid declines of native European ladybirds. *Diversity and Distributions* 18, 7 (2012), 717–725. DOI:http://dx.doi.org/10.1111/j.1472-4642.2012.00883.x

Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proc. the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 614–622.

Aashish Sheshadri and Matthew Lease. 2013. SQUARE: A benchmark for research on computing crowd consensus. In *First AAAI Conference on Human Computation and Crowdsourcing*.

Jonathan Silvertown, Martin Harvey, Richard Greenwood, Mike Dodd, Jon Rosewell, Tony Rebelo, Janice Ansine, and Kevin McConway. 2015. Crowdsourcing the identification of organisms: A case-study of iSpot. *ZooKeys* 480 (2015), 125.

Jeffrey S Simonoff. 1995. Smoothing categorical data. *Journal of Statistical Planning and Inference* 47, 1 (1995), 41–69.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proc. the conference on empirical methods in Natural Language Processing*. Association for Computational Linguistics, 254–263.

Wei Tang and Matthew Lease. 2011. Semi-supervised consensus labeling for crowdsourcing. In *SIGIR 2011 workshop on crowdsourcing for information retrieval (CIR)*.

Matteo Venanzi, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi. 2014. Community-based bayesian aggregation models for crowdsourcing. In *Proc. the 23rd international conference on World wide web*. International World Wide Web Conferences Steering Committee, 155–164.

Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier R Movellan, and Paul L Ruvolo. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems*. 2035–2043.

Yan Yan, Glenn M Fung, Rómer Rosales, and Jennifer G Dy. 2011. Active learning from crowds. In *Proc. the 28th international conference on machine learning (ICML-11)*. 1161–1168.

### A. BAYES RULE IN ODDS FORM

Bayes rule allows the conditional probability of a hypothesis $H$ (or proposition) being true given evidence $E$ has been observed, $P(H|E)$, to be rewritten in terms of the conditional probability of seeing the evidence for a given hypothesis, $P(E|H)$, and the individual probabilities of the hypothesis, $P(H)$, and evidence, $P(E)$:

$$P(H|E) = P(E|H) \times \frac{P(H)}{P(E)} \tag{9}$$

Bayes rule follows from the definition of joint probabilities:

$$P(H, E) = P(H|E) \times P(E) \tag{10}$$

That is, the joint probability of $H$ being true and $E$ being observed is the probability of $E$ being observed multiplied by the conditional probability of $H$ being true, given $E$ has been observed. Equivalently, the joint probability of $E$ being observed and $H$ being true is the probability of $H$ being true multiplied by the conditional probability of $E$ being observed, given $H$ is true:

$$P(E, H) = P(E|H) \times P(H) \tag{11}$$

As joint probabilities are reflexive, $P(H, E) = P(E, H)$ and the right hand sides of (10) and (11) are the same, resulting in (9), known as Bayes rule. Similarly. the conditional probability of the hypothesis $H$ being false given the evidence $E$ is observed is:

$$P(\neg H|E) = P(E|\neg H) \times \frac{P(\neg H)}{P(E)} \tag{12}$$

Dividing (9) by (12), we get Bayes rule in odds form:

$$\frac{P(H|E)}{P(\neg H|E)} = \frac{P(H)}{P(\neg H)} \times \frac{P(E|H)}{P(E|\neg H)} \tag{13}$$

This is often written as:

$$O(H|E) = O(H) \times \Lambda(H|E) \tag{14}$$

where the a priori odds of a hypothesis being true $O(H)$ are defined as the ratio of the a priori probability of the hypothesis being true, $P(H)$, to being false, $P(\neg H)$, and the lambda term ($\Lambda$), the ratio of the conditional probability of seeing the evidence for the hypothesis, $P(E|H)$ to seeing the evidence for any other hypothesis $P(E|\neg H)$ is given by:

$$\Lambda(H|E) = \frac{P(E|H)}{P(E|\neg H)} \tag{15}$$

Bayes rule is so powerful because the quantities on the right hand side of (14) can be directly estimated from data, as we do in this paper. This allows us to update the odds of a hypothesis being correct based on new evidence by a user. If there are multiple observations of evidence $E_1$ to $E_n$,

$$O(H|E_1, E_2, ..., E_n) = O(H) \times \Lambda(H|E_1, E_2, ..., E_n) \tag{16}$$

$$\text{where, } \Lambda(H|E_1, E_2, ..., E_n) = \frac{P(E_1, E_2, ..., E_n|H)}{P(E_1, E_2, ..., E_n|\neg H)} \tag{17}$$

If each observation of evidence is independent, as it would be in biological recording when different users identify the specimen independently, the joint probabilities in the $\Lambda$ term in (17) are just the products of the individual probabilities:

$$P(E_1, E_2, ..., E_n|H) = P(E_1|H) \times P(E_2|H) \times ... \times P(E_n|H) \tag{18}$$

$$P(E_1, E_2, ..., E_n|\neg H) = P(E_1|\neg H) \times P(E_2|\neg H) \times ... \times P(E_n|\neg H) \tag{19}$$

Dividing (18) by (19), we see that:

$$\Lambda(H|E_1, E_2..., E_n) = \Lambda(H|E_1) \times \Lambda(H|E_2) \times ... \times \Lambda(H|E_n) \tag{20}$$

Substituting (20) in (16), it can be seen that for independent identifications by users, the odds for $H$ being true given observation of $E_1...E_n$ can be computed just by multiplying the prior odds $O(H)$ with the $\Lambda$ terms for individual observations of evidence:

$$O(H|E_1, E_2, ..., E_n) = O(H) \times \Lambda(H|E_1) \times \Lambda(H|E_2) \times ... \times \Lambda(H|E_n) \tag{21}$$

Given the definition in (21), Bayes rule in odds form can be used incrementally to re-evaluate the probability of a hypothesis each time new evidence comes in, just by multiplying the existing odds with the $\Lambda$ term corresponding to the new evidence:

$$O(H = S_j|E_1, ..., E_{n+1} = S_i) = O(H = S_j|E_1, ..., E_n) \times \Lambda(H = S_j|E_{n+1} = S_i) \tag{22}$$

This is the form in which Bayes rule is implemented in the work described in this paper. Odds can be converted to probabilities and vice versa using the formulas below:

$$O(X) = \frac{P(X)}{1 - P(X)} \tag{23}$$

$$P(X) = \frac{O(X)}{1 + O(X)} \tag{24}$$